Ego-PMOVE: Prompt-aware Mixture of View Experts Network for Egocentric Gaze Prediction

Heqian Qiu, Lanxiao Wang*, Taijin Zhao, Zhaofeng Shi, Xiang Li, Linfeng Xu, Hongliang Li*

University of Electronic Science and Technology of China, Chengdu, China
{hqqiu,lanxiaowang}@uestc.edu.cn, {zhtjww,zfshi,202411012315}@std.uestc.edu.cn, {lfxu,hlli}@uestc.edu.cn

Abstract

Egocentric gaze prediction serves as a critical indicator for decoding human visual attention and cognitive processes, but its inherently limited field of view creates prediction challenges. Although exo-view data provides supplementary contextual information, it exhibits significant spatial and semantic gaps. Existing methods focus solely on isolated feature encoding in single-view paradigms, neglecting cross-view gaze correlations. To make up for this gap, we make the first exploration of cross-view gaze relationship for egocentric gaze prediction, and propose Ego-PMOVE, a novel Prompt-aware Mixture of View Experts network. Unlike prior cross-view studies that forcibly align cross-view features thereby introducing inference noise, we leverage the popular Mixture-of-Experts (MoE) and a set of flexible prompts to disentangle features from different views into three parallel experts: a view-shared expert directly modeling common semantic relationships, a view-discrepancy expert adaptively adjusting the spatial position, scale and shifts based on different view-specific features, and an egocentric expert extracting independent features to compensate for the case of missing exocentric data. To balance these experts, we further design a soft router to dynamically weight them for mining useful information while suppressing noise. A view-query gaze decoder then generates view-specific gaze attention maps, jointly optimized by gaze-heamap and cross-view contrastive loss that regularize both shared and divergent features for accurate gaze prediction. Extensive experiments across the multi-view EgoMe dataset and single-view Ego4D and EGTEA Gaze++ datasets demonstrate the effectiveness and generalizability of our approach.

Code — https://github.com/QiuHeqian/Ego-PMOVE

Introduction

With the increasing prevalence of wearable cameras, egocentric vision that emulates human perception has emerged as a critical frontier in computer vision, focusing on parsing videos captured from a first-person perspective. As egocentric eye movements provide the important visual attention information about the camera wearer, predicting egocentric visual attention, known as egocentric gaze prediction, is thus a key step towards understanding and inferring the human behavioral intent and cognitive processes. This capability enables valuable applications in Augmented Reality (AR), Virtual Reality (VR), and embodied AI systems(Shi, Dao, and Cai 2025).

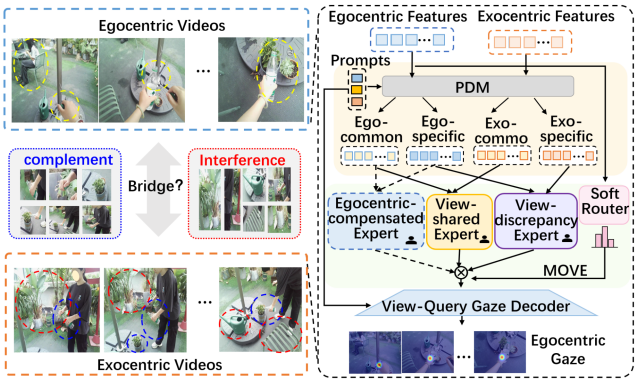

Figure 1: An example of cross-view gaze data and the main idea of our proposed Ego-PMOVE method. the red, blue and yellow dotted boxes are interference noises, complementated information, and challenge regions.

Notably, egocentric vision scenes inherently suffer from a limited field-of-view, object occlusion, and dynamic motion blur, making egocentric gaze prediction particularly challenging. A natural remedy is to introduce other-view videos, which provide global context and spatial cues complementary to the wearer's narrow perspective, as shown in Fig. 1. However, current egocentric gaze prediction methods (Li, Liu, and Rehg 2018; Lai et al. 2024a,b) remain confined to single-view encoding paradigms. They exclusively extract egocentric video features to estimate the wearer's attention distribution, while neglecting to model cross-view gaze behavior correlations for complementary cues. Although recent studies have leveraged exocentric videos for tasks such as video captioning (Xu et al. 2024; Grauman et al. 2024), its potential for improving egocentric gaze prediction in multi-view settings-crucial for decoding human intent and cognitive process—has yet to be explored.

This gap persists because existing cross-view techniques are difficult to ill-suited for fine-grained gaze prediction.

*Corresponding authors.

They typically enforce strict alignment by forcibly mapping cross-view egocentric and exocentric features into a shared latent space (Xue and Grauman 2023; Ohkawa et al. 2023). As shown in Fig. 1, despite valuable complementary information existing between egocentric and exocentric views, e.g., the "bottle", the human global "pour" action and "flowerpot", significant discrepancies persist in viewpoint-induced spatial shifts and semantic granularity gaps. Such strict alignment overlooks the essence of egocentric gaze prediction, which not only easily introduces noise, but also erases the distinctive dynamics unique to each perspective. For example, in the upper right corner of Fig. 1's final frame, egocentric video contains "bottle" and "flowerpot" objects whereas exocentric video represents a human body. In addition, objects appear significantly larger in egocentric compared to exocentric video. These cross-view divergences impede direct integration, compelling the key challenge of how to effectively exploit complementary cross-view information and bridge these view gaps for accurate gaze prediction.

To address the above challenges, we make the first cross-view exploration for gaze prediction, propose a novel prompt-aware view-mixture-of-experts network (Ego-PMOVE) to adaptively bridge the gap between exo- and ego-view information. Leveraging the powerful multi-task gating of Mixture-of-Experts (MoE) (Jacobs et al. 1991; Pavlitska et al. 2023) and the flexible representation of visual prompts (Radford et al. 2021), we explicitly feed different cross-view features into three collaborative expert pathways for avoiding feature interference between different characteristics. Specifically, we first design a prompt-aware disentangle module, which uses a set of learnable prompts including view-shared prompts and egocentric/exocentric-specific prompts to flexibly encode common semantics and unique view information respectively. Accordingly, the input features from different views can be dynamically disentangled into view-common and view-specific features that remain isolated, providing a clear and uncontaminated representation for subsequent divide-and-conquer expert processing. Building on these disentangled features, moreover, we propose a mixture-of-view-experts model with three parallel pathways: (1) a view-shared expert that naturally builds bidirectional cross-attention interaction between the disentangled exocentric-common and egocentric-common features; (2) a view-discrepancy expert adaptively adjusts feature saptial positions and semantic shifts for eliminating view gap based on a learnable deformation component and a semantic modulation module, effectively achieving view features alignment and complement. (3) an egocentric-compensated expert operates independently on the egocentric features, leveraging self-attention and several linear layers to compensate for any missing exocentric information.

To adaptively balance these experts, we further design a soft router to adaptively assign expert weights based on the input video's viewpoint completeness. Finally, the fused features are fed into a view-query gaze decoder, which can generates egocentric gaze attention heatmaps via cross-view multi-scale queries and supervises them with the corresponding gaze distributions. A cross-view contrastive loss is designed to pull together shared common features from the same viewpoint while pushing them away from view-specific features, ensuring a clear separation between common semantics and viewpoint-private dynamics for accurate gaze prediction. We conduct extensive experiments on multi-view EgoMe and single-view datasets to demonstrate the effectiveness of our method.

The main contributions are summarized as follows:

- We present the first exploration to bridge the gap in cross-view gaze correlation for egocentric gaze prediction.
- We propose a novel prompt-aware view-mixture-of-experts network (Ego-PMOVE), including prompt-aware disentangle module, mixture of view experts, view-query gaze decoder and cross-view contrastive loss.
- Our method significantly outperforms existing state-of-the-art methods on the multi-view EgoMe dataset and single-view Ego4D and EGTEA Gaze++ datasets, which demonstrates the effectiveness and generality even without exocentric gaze data.

Related Works

Egocentric Gaze Prediction. Due to the fact that egocentric videos more naturally reflect human intrinsic perception, a number of studies have emerged focused on the egocentric gaze prediction task. (Li, Fathi, and Rehg 2013) proposed a graphical model by relying predefined egocentric eye-hand/head coordinate cues and the temporal dynamics of gazes, enabling predict gaze position at each frame. Later, they jointly learned gaze and actions by modeling the gaze distribution using stochasitic units in (Li, Liu, and Rehg 2018). (Zhang et al. 2017) developed a generative adversarial neural network with a two-stream 3D-CNN generator to anticipate gaze location on future frames. (Huang et al. 2018, 2020) integrated a task-dependent attention transition with LSTM and bottom-up visual saliency model to explore the gaze attention region based on previous fixations. (Tavakoli et al. 2019) combined bottom-up and top-down attentional cues to guide egocentric gaze prediction. (Thakur et al. 2021) investigated the gaze prediction improvement achieved through the joint fusion of IMU data, first-person optical flow and RGB data. (Lai et al. 2024a) designed a transformer-based global-local correlation model to explicitly capture the relationship between global context and each local token. Furthermore, they introduced audio information in addition to original videos, and proposed a spatial-temporal separable fusion method (Lai et al. 2024b) to separately capture audio-visual correlations in spatial and temporal dimensions. However, these existing methods remain narrowly capture egocentric gaze from a single-view demonstration, thereby failing to associate and integrate cross-view gaze information between egocentric and exocentric videos.

Cross-view Video Understanding. It is necessary to investigate the visual associations between different views, as they provide diverse complementary information, especially exo-ego data. Recently, a series of exo-ego video datasets have emerged to further promote the development of cross-view video understanding (Sigurdsson et al. 2018; Huang et al. 2024; Grauman et al. 2024; Shi, Hayat, and Cai 2024). They are typically collected using eye-tracking cameras for egocentric-view video recording, along with multi-

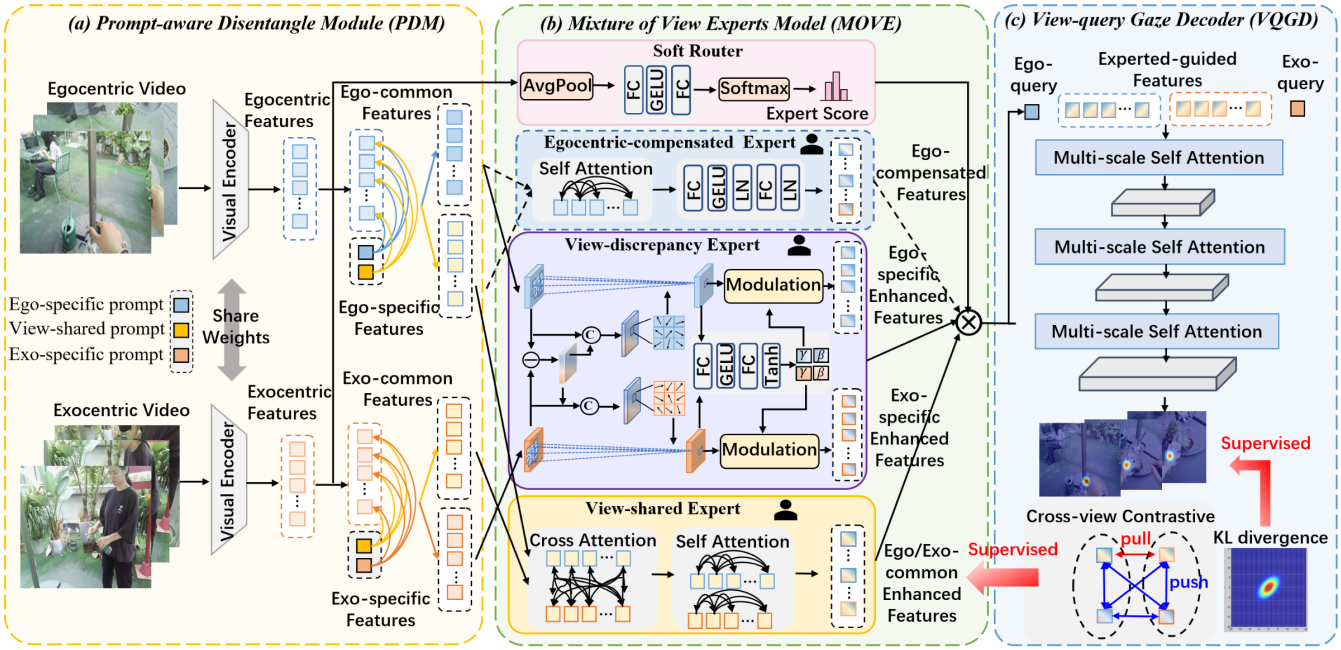

Figure 2: The overall network architecture of our proposed Ego-PMOVE for egocentric gaze prediction, including prompt-aware disentangle module (PDM), mixture of view experts model (MOVE), view-query gaze decoder (VQGD).

ple fixed cameras to capture exocentric-view videos. In addition, some researchers (Wang et al. 2023; Park, Lee, and Sohn 2025; Xue and Grauman 2023; Li et al. 2021; Jiang et al. 2022, 2023, 2019; Qiu et al. 2024) attempt to learn vie-invariant video representations with the contrastive learning method or transfer cross-view feature representations via knowledge distillation. Meanwhile, more fine-grained exo-ego video generation (Liu et al. 2021, 2024) has been tried to explored based on generative adversarial network or diffusion model. Building upon this, recent efforts have been made to leverage cross-view visual information for augmenting egocentric video understanding, including action recognition (Yu et al. 2019; Rahmani and Mian 2015; Qiu et al. 2022), egocentric video caption (Xu et al. 2024; Ohkawa et al. 2023), sequence verification (Li et al. 2024), action anticipation (Huang et al. 2024). Thanks to the innovative proposal of the EgoMe dataset (Qiu et al. 2025), which provides information on exocentric observing gaze and egocentric following gaze information. In this paper, we systematically explore the cross-view gaze relationship for delve deeper into revealing human attention processes.

The Proposed Method

In this paper, we propose a novel prompt-aware mixture of view experts network (Ego-PMOVE) that adaptively bridges cross-view correlations to assist in limited-of-field egocentric gaze prediction. The overall architecture is illustrated in Fig. 2. Given a pair of egocentric and exocentric videos, we first adopt a popular CLIP-ViT encoder to extract their video features. Then, we develop a prompt-aware disentangle module to flexibly disentangle the original encoded features into ego/exo-common, ego/exo-specific features. Based on these disentangled features, we further design a mixture of view experts model with three parallel expert branches, which separately models cross-view correlations for common cross-view features, view-specific features and egocentric-only features. Meanwhile, we utilize a soft router to dynamically balance these experts, preventing mutual interference among heterogeneous features. Finally, A view-query gaze decoder utilizes the view-specific prompts to gradually query and decode the corresponding gaze attention, while a gaze attention distribution loss combined with a cross-view shared-discrepancy contrastive loss jointly optimizes the network and feature distances.

Prompt-aware Disentangle Module

Recently, prompt has been demonstrated the powerful flexible encoding capabilities in mutimodal tasks (Radford et al. 2021). Here, we design a prompt-aware disentangle module that leverages a set of learnable prompts to explicitly extract and decouple different characteristics from exocentric and egocentric features. Given a pair of exocentric and egocentric videos V_{exo}, V_{ego} with T frames, we first adopt the shared-weight CLIP-ViT as the backbone to extract visual features $F_{exo}, F_{ego} \in \mathcal{R}^{THW \times C}$ for corresponding-view videos, where H, W and C denotes the height, width and channel of the visual features. Let a learned prompt vectors set $\mathcal{P} = \{P_c, P_{ego}, P_{exo}\} \in \mathcal{R}^{3 \times C}$, which encodes cross-view shared semantic representation, view-specific features for egocentric and exocentric views, respectively. Then, we perform feature disentanglement by computing cross-attention between these

prompts and visual features, which takes the visual features as query, the prompts as the key and value. F^c_{exo}, F^c_{exo} denotes the shared exocentric/egocentric-common features extracted by the view-shared prompts. F^s_{exo}, F^s_{exo} separately denotes the exocentric/egocentric-specific features extracted by the prompts P_{ego}, P_{exo}.

Mixture of View Experts Model

To overcome the limitations of previous strict cross-view alignment, we propose a mixture of view experts model (MOVE) that achieves cross-view feature fusion guided by a "divide-and-conquer" principle through three specialized fusion experts. The overall module consists of view-shared expert, view-discrepancy expert, egocentric-compensated expert and a soft router. These experts are specifically designed according to the characteristics of the input view features. Meanwhile, the soft router is employed to dynamically balance the contributions of different experts. This design allows the network to flexibly leverage value semantic information for comprehensive feature complement while preserving view-exclusive information and avoiding interference from irrelevant noise.

View-shared Expert. Based on the above decoupled exo/ego-common features, we introduce a view-shared expert to directly model spatial-temporal relationships for view-invariant feature fusion. Due to the feature sharing across perspectives, we don't need to perform any specific perspective transformation. Instead, we compute the relationships between exo-common and ego-common features using two cross-attention mechanism. Specifically, we take the exo-common visual features as query, the ego-common features as the key and value via the linear layers, and reweight the features from ego-common perspective after being processed through a softmax function, otherwise. Additionally, we integrate a semantic self-attention layer to facilitate internal interaction within each view feature, ultimately optimizing the fused feature representation. $F^s e_{exo}, F^s e_{exo}$ separately denotes the exocentric/egocentric-common enhanced features by the view-shared expert.

View-discrepancy Expert. For ego/exo-specific features with significant spatial and semantic differences, we design a view-specific expert. This expert explicitly adjusts spatial position offsets via a discrepancy-guided deformable convolution module and performs feature scaling/shifts with a feature modulation module to align cross-view features for effective fusion.

Specifically, we first calculate the feature difference F_d between the exocentric F^s_{exo} and egocentric F^s_{exo} views, and then concatenate the difference with view-specific features to generate difference-aware offset vectors $\Delta O_{exo}, \Delta O_{ego}$ using two convolution layers with $k \times k$ kernel size:

$$\begin{aligned} F_d &= |F^s_{exo} - F^s_{ego}|, \\ F^d_{exo,ego} = [F^s_{exo,ego}; F_d], &\Delta O_{exo,ego} = W_d F^d_{exo,ego}, \end{aligned} \tag{1}$$

where $|\cdot|$ indicates the absolute value operator, W_d are weight parameters. $\Delta O^{exo,ego} \in \mathcal{R}^{THW \times 2k^2}$ separately denote the spatial position offsets for the features F^s_{exo} and F^s_{ego}. For simplicity, we take the egocentric view as an example and omit the subscript for the following processing.

Given the p-th position feature $F^s(p)$, $\Delta p = \{\Delta p_1, ...\Delta p_i, ...\Delta p_{k^2}\} \in \mathcal{R}^{2k^2}$ is the p-th position spatial offsets of ΔO_{exo}, we can add the difference-aware offsets to adjust the initial alignment location and then weighted aggregated these spatial points to generate the aligned features $F^s_a(p)$ at each position p for egocentric or exocentric view:

$$F^s_a(p) = \sum_{i=1}^{k^2} W_i \cdot F^s(p + p_i + \Delta p_i), \tag{2}$$

where W_i, p_i, Δp_i is the i-th kernel weight, the i-th initial offsets and compensated offsets, respectively. Hence, $p_i \in \{(-1,-1),(-1,0),...,(1,1)\}$ denotes a regular grid of a 3×3 kernel defined with $k = 3$, in which each set of coordinates represents the relative position of a point to p. $F^s_a(p)$ represents the aligned features based on the cross-view difference features.

Based on spatial-aligned features $F^s_a(p)$, we further design a feature modulation module to adaptively align cross-view semantic by applying an affine transformation. We first utilize a feed-forward neural network (FFN) including two linear layers and a GeLU activate function to adaptively generate the scale and shift values, and then use a tanh function to map the value into the range of $[-1, 1]$. Finally, we adopt the mapped scale γ and shift β to modulate the per-feature-map distribution via a feature-wise affine transformation, agnostic to spatial location:

$$\begin{aligned} [\gamma, \beta] &= tanh(FFN(F^s_a)), \\ F^m_a &= \gamma \odot FFN(F^s_a) + \beta, \end{aligned} \tag{3}$$

where $\odot$ denotes scalar multiplication, F^m_a represents the modulated features after scaling them up or down.

Egocentric-compensated Expert. Because it is not always available for exocentric gaze maps, we introduce an egocentric-compensated expert that automatically generates surrogate representations of exocentric gaze, which not only effectively compensates for missing auxiliary information but also significantly enhances model robustness in real-world applications. Given the egocentric features $F_{ego} \in \mathcal{R}^{THW \times C}$ extracted by CLIP, we first deploy a self-attention Layer to model the spatio-temporal dependencies within the egocentric gaze features and capture long-range interaction in key regions. Subsequently, we adopt a FFN composed of two Linear layers, LayerNorm, and a GeLU activation function to project the enhanced egocentric features into a shared embedding space $F^e \in \mathcal{R}^{THW \times 2C}$, which can be split into $F^e_{ego} \in \mathcal{R}^{THW \times C}$ for egocentric-compensated features and $F^e_{exo} \in \mathcal{R}^{THW \times C}$ for compensating exocentric features.

Soft Router. After obtaining the fused features from different experts, an assignment mechanism is needed to better coordinate each expert. Unlike previous hard MOE that only allows a single expert to access the input, our task acknowledges that each video may contain both shared semantics and distinct features from different experts. To address this complexity, we design a soft router to allow multiple experts to process the input simultaneously. Specifically, given the

pair of egocentric and exocentric features F_{ego} and F_{exo}, we first concatenate their features and then employ a gating network to perform global average pooling, a FFN and a softmax function to generate scores for each pair of cross-view video features corresponding to each expert. This design enables the model to dynamically assign input data to multiple experts based on their relevance, thereby effectively handling the intricate nature of the input information:

$$g = Softmax(FFN(AvgPool([F_{exo}; F_{ego}]))), \quad (4)$$

$$F' = F + \sum_{i=1}^{3} g_i \cdot F^i, \quad (5)$$

where $g = \{g_1, g_2, g_3\} \in \mathcal{R}^3$, $g_i \in g$ is the gate control vectors and importance for the i-th expert. The soft mechanism is built on the fact that the input cross-view features can adaptively determine how much (weight) should be sent to each expert by the softmax function. The larger the value of g_i, the more important the feature F^i is considered to be. F^i and F denote the fused features output by the i-th expert and balanced features cooperated among different experts.

View-query Gaze Decoder

To generate gaze map distributions corresponding to different views, we design a view-query gaze decoder to gradually decode gaze attention map with the desired resolution and view. Given the view-specific prompts P_{ego}, P_{exo} and fused multi-view features F'_{ego}, F'_{exo} from MOVE, we first concatenate corresponding view features and prompts $[P_{ego}; F'_{ego}]$ and $[P_{ego}; F'_{ego}]$. Then, we feed the concatenated view features into its corresponding transformer decoder based on the multi-scale self-attention block of MViT (Fan et al. 2021) to progressively decode the gaze attention maps up to the desired resolution for exocentric and egocentric views. Within each decoder block, following (Lai et al. 2024a), we upsample the input features with prompt as query Q_i, while pooled features serve as keys and values K_i and V_i. The self-attention layer then computes correlation weights between Q_i and K_i, aggregates the values V_i accordingly, and outputs decoded features F^D_{i+1}. These decoded features are further fused with the original input features via skip connections to generate decoded gaze feature representations.

$$F^D_{i+1} = Softmax(Up(Q_i)Pool(K_i)/\sqrt{C})Pool(V_i) + F^D_i, \quad (6)$$

Here, both feature upsampling Up and pooling $Pool$ operations are implemented using trilinear interpolation. Our decoder consists of 4 decoding blocks. In each block, the spatial resolution of features is progressively increased (e.g., from lower to higher resolutions). Finally, after processing through all decoding blocks, a convolutional layer transforms the highest-resolution features to produce the final gaze map prediction.

Network Optimization

In this paper, we propose an end-to-end optimized network training strategy supervised by a combination of Kullback-Leibler (KL) divergence loss and cross-view contrastive loss. On the one hand, the KL divergence loss measures the distributional differences between the predicted gaze maps for ego- and exo- views and their corresponding ground-truth labels that are converted into heatmaps by centering a 2D Gaussian distribution at the gaze location. On the other hand, to effectively discover shared and discrepancy features across views, we design a cross a cross-view contrastive loss, which directly supervises the fused features of different experts via a contrastive learning mechanism. Different from traditional contrastive learning, our approach encourages the shared features $F^c_{exo,ego}$ extracted by the shared expert using global average pooling from different views to be as close as possible, while pushing the shared and distinctive features within the same view $F^c_{exo,ego}, F^s_{exo,ego}$, as well as the distinctive features $F^s_{exo,ego}$ across different perspectives, to be as far apart as possible. The overall loss is computed below:

$$\mathcal{L} = \sum_{i \in exo,ego} \mathcal{L}_{kl}(G^p_i, G^*_i) + \lambda \mathcal{L}_{cvc}(F^c_{exo,ego}, F^s_{exo,ego}),$$

$$\mathcal{L}_{cvc} = \frac{exp(F^c_{exo} F^c_{ego})}{\sum_{i=c,s} exp(F^i_{exo} F^i_{ego}) + \sum_{i=exo,ego} exp(F^c_i F^s_i)} \quad (7)$$

where λ is used to control the contributions for the KL divergence loss $\mathcal{L}_{KL}$ and the cross-view contrastive loss $\mathcal{L}_{cvc}$.

Experiments

Experimental Setup

Datasets and Evaluation Metrics. To verify the effectiveness of our method comprehensively, we conduct our experiments on both multi-view EgoMe dataset and two sing-view EGTEA Gaze++ and Ego4D egocentric gaze datasets. Specifically, the EgoMe dataset (Qiu et al. 2025) spans 41 daily-life scenes, provides 6148 paired cross-view videos with exocentric gaze recordings from observers and egocentric gaze videos from imitators performing the same actions. To ensure cross-view action consistency, we clip original long videos by sub-actions, yielding 3,706 cross-view gaze pairs (16,650 clips) for training, 799 pairs (3,647 clips) for validation, and 1,643 pairs (7,353 clips) for testing. The EGTEA Gaze++ dataset (Li, Liu, and Rehg 2018) is collected in natural kitchen environments, containing 8,299 video splits for training and 2,022 splits for testing. The Ego4D (Grauman et al. 2022) offers 27 videos totaling 31 hours of gaze-tracking data under the social setting. Following common settings (Lai et al. 2024a), we split the long videos into 5-second video clips with gaze fixation, allocating 20 videos (15,310 clips) for training and 7 videos (5,202 clips) for testing. Similar to recent works, we adopt F1 score (primary), recall and precision as gaze evaluation metrics.

Implemental Details. In our experiments, we adopt CLIP with ViT-B/32 to initiate egocentric and exocentric video visual encoders. we sample 8 frames for each video resized to 224x224 spatial resolution. The dimensions of channel embedding are sets as $C = 768$, the view-contrastive learning loss weight is set to $\lambda = 0.1$. We employ the AdamW optimizer for 10 epochs with an initial learning rate of 3e-4, weight decay 0.05, momentum 0.9, and total batch size 16 distributed across 4 GPUs.

Exo	PDM	MOVE	VQGD	F1	Recall	Precision
				49.84	60.92	42.17
✓				50.06	61.26	42.33
✓		✓		52.08	61.84	44.98
✓	✓	✓		52.45	62.80	45.03
✓			✓	51.86	62.35	44.39
✓	✓		✓	52.93	62.10	46.12
✓		✓	✓	52.95	62.93	45.70
✓	✓	✓	✓	**53.44**	**63.01**	**46.39**

Table 1: The effects of main components in our method.

Method	F1	Recall	Precision
Ego-only	49.84	60.92	42.17
co-training	50.06	61.26	42.33
Vanilla Linear	50.15	61.65	42.27
Self-attention	50.40	61.46	42.72
w/o ECE	52.13	62.58	44.68
w/o VSE	51.68	62.30	44.16
w/o VDE	51.49	62.27	43.89
w/o SoftRouter	51.74	62.60	44.09
MOVE	**52.45**	**62.80**	**45.03**

Table 2: The effects of MOVE module in our method.

Method	F1	Recall	Precision
w/o VQGD	50.06	61.26	42.33
SGD	51.93	61.97	44.69
MGD	52.60	63.32	44.99
VQGD	**52.93**	**62.10**	**46.12**

Table 3: The effects of VQGD module in our method.

Method	F1	Recall	Precision
KL	52.66	62.64	45.42
KL+CSL	52.94	63.27	45.51
KL+CDL	52.97	63.65	45.36
KL+CCL	**53.44**	**63.01**	**46.39**

Table 4: The effects of CCL mechanism in our method.

Ablation Studies

Here, we conduct ablation studies to verify the effectiveness of our main components on EgoMe validation set.

Effects of Main Components. Table 1 shows the effects of main components in our method Ego-PMOVE. For fair comparison, we adopt the same CLIP-ViT model trained on egocentric videos as our baseline. Compared to the baseline model, introducing exocentric videos as co-training data yields only a slight improvement, indicating the necessity of effectively utilizing information from exocentric videos as complement. When integrating our proposed MOVE model, the performance significantly increased by 2.24% in terms of F1 score, demonstrating its effectiveness. Furthermore, the proposed VQGD model contributes a 2.02% F1 score improvement over the baseline, validating its importance for gaze map prediction. The simultaneous adoption of MOVE and VQGD achieves an overall performance of 52.95%, representing incremental improvements of 0.87% and 1.09% over using either module alone. Moreover, when we introduce the PDM for feature disentanglement on the above methods, further performance gains were observed, proving that decoupled features facilitate model processing. Ultimately, our overall framework reaches 53.44% F1 score, significantly surpassing the baseline method by 3.6%.

Effects of Mixture of View Experts Model. We analyze the impact of our proposed Multi-view Expert (MOVE) model in Table 2. The upper table compares performance across different cross-view fusion strategies: 1) Co-training: training the model from both egocentric and exocentric gaze data. 2) Vanilla Linear: Fusing features by a few of linear layers. 3) Self-attention: Fusing cross-view vision tokens via standard self-attention mechanisms. It can be observed that these methods all show limited performance improvements compared to the ego-only baseline, indicating that these methods are suboptimal for cross-view feature fusion. In contrast, our proposed method demonstrates a larger performance boost by 2.61%. In addition, the lower table shows the contribution of different experts: egocentric-compencated expert (VCE), view-shared expert (VSE), view-discrepancy expert (VDE). It can be seen that removing any expert leads to a decline in model performance, proving the importance of each expert in our method. The VDE expert has a relatively larger impact, with a decrease of 0.96%, while the ECE expert has a smaller impact, with a decrease of 0.32%, since it focuses only on egocentric video information. Moreover, the soft router can improve performance by 0.71%, demonstrating its effectiveness in balancing the contributions of different experts.

Effects of View-query Gaze Decoder. Table 3 shows the performance impact of the proposed View Query Guided Decoder (VQGD). It can be observed that without view queries, the single-scale gaze decoder (SGD) yields a 1.89% performance improvement over the simple upsampling baseline (w/o VQGD), demonstrating the effectiveness of the proposed decoding module. Furthermore, the multi-scale gaze decoder (MGD) achieves 0.67% gain over the single-scale version. When view queries are introduced, the performance reaches its optimal level, indicating that the proposed VQGD are beneficial for accurate egocentric gaze decoding.

Effects of Cross-view Contrastive Loss. Table 4 analyzes the impact of different contrastive loss settings. When supervised solely by the KL divergence loss for gaze heatmap, the model achieves 52.66% performance. Adding either the cross-view shared loss (CSL, minimizing cross-view shared feature distances) or cross-view discrepancy loss (CDL, maximizing distinctive-feature distances) individually yields marginal gains of approximately 0.3%. However, when we employ the cross-view contrastive loss to constrain both types of features simultaneously, the performance improves by 0.78%, demonstrating that the proposed feature loss is effective for cross-view fusion.

drop rate	0	0.1	0.3	0.5	0.8	1.0
F1	53.44	53.54	53.26	53.51	53.45	52.61

Table 5: The effects of drop rate of exocentric video.

Comparison with State-of-the-art Methods

We compare our proposed Ego-PMOVE method with state-of-the-art methods across three benchmark datasets: a cross-view gaze dataset EgoMe, and two single-view datasets Ego4D and EGTEA Gaze++. In Table 6, for a comprehensive comparison, we not only re-implemented existing methods (GLC (Lai et al. 2024a) and CSTS (Lai et al. 2024b)) on EgoMe datasets but also employed several typical backbone networks (such as I3D-R50 (Feichtenhofer et al. 2019), MViT (Fan et al. 2021), and CLIP-ViT (Radford et al. 2021)) for the egocentric gaze prediction task on the EgoMe (Qiu et al. 2025) dataset. These results indicate that our method significantly outperforms EgoMe* with cross-view gaze data by 6.12% and 6.08% on the validation and test sets, respectively. Compared to the previous best existing method CSTS, our method also consistently achieve higher performance by 4.03% and 4.66% in terms of F1 score. For single-view gaze dataset in 7, our method still surpasses the best method CSTS, by 2.82% on Ego4D and outperforms previous best model GLC on the EGTEA Gaze++ dataset by 3.59%. In addition, we analyze the effects of drop rate of exocentric videos in Table 5, there is almost no change in performance here when we random remove exocentric videos. We set 0.3 occlusion, F1 is almost unchanged(53.22 vs 53.44). These results demonstrate our method's effectiveness in cross-view gaze data and its strong generalization capability to single-view gaze scenarios without exocentric supervision.

Visualization

Fig. 3 visualizes successful (left) egocentric gaze map prediction results of our Ego-PMOVE model and the baseline method. It can be observed that for [specific case description, e.g., complex interactive scenes], the baseline that only use egocentric video fails to localize the correct gaze target and may be misled by other salient interfering objects. In contrast, our proposed method achieves accurate gaze prediction by effectively leveraging complementary information from the exocentric view.

Conclusion

In this paper, we propose a novel Prompt-aware Mixture of View Experts Network (Ego-PMOVE) for egocentric gaze prediction. Based on MOE and flexible prompts, we first disentangle features from different views into three parallel experts: a view-shared expert directly for common semantic relationships, a view-discrepancy expert adaptively adjust the spatial position, scale and shifts based on view-difference features for view-specific features, and a egocentric expert for compensating missing exocentric data. To balance these experts, we further design a soft router to dynamically weight the experts for mining useful information while suppressing noise. A view-query gaze decoder generates view-specific attention maps, optimized jointly by gaze-distribution and cross-view contrastive losses that regularize both shared and divergent features for accurate gaze prediction. Experimental results on multi-view EgoMe dataset and the single-view Ego4D and EGTEA Gaze++ datasets demonstrate the effectiveness and generalizability.

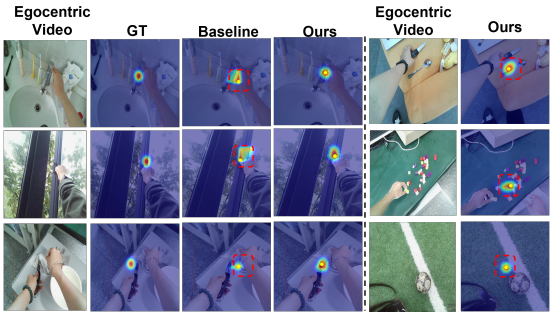

Figure 3: The successful (left) and failure (right) visualization egocentric gaze map prediction results. The red dotted boxes represent false gaze prediction.

Method	Val			Test		
	F1	Rec	Prec	F1	Rec	Prec
I3D-R50	43.38	57.16	34.95	42.81	56.47	34.47
MViT	47.74	59.66	39.79	48.44	61.10	40.13
CLIP-ViT	49.84	60.92	42.17	49.98	60.51	42.57
GLC	48.83	60.69	40.85	48.95	59.41	41.62
CSTS	49.41	59.75	42.12	49.32	62.84	40.59
EgoMe	47.11	60.40	38.62	47.66	66.09	49.91
EgoMe*	47.32	59.67	39.20	47.90	60.20	39.77
Ours	**53.44**	**63.01**	**46.39**	**53.98**	**62.19**	**47.68**

Table 6: Comparison with state-of-the-art methods on EgoMe dataset. EgoMe* denotes co-training using cross-view data. To save space, we use abbreviations: Rec for Recall, Prec for Precision.

Method	Ego4D			EGTEA Gaze++		
	F1	Rec	Prec	F1	Rec	Prec
Center Prior	14.9	21.9	11.3	10.7	32.0	6.4
GBVS	18.0	47.2	11.1	15.7	45.1	9.5
GazeMLE	35.4	49.7	27.5	26.6	35.7	21.3
AttnTrans	36.4	47.6	29.5	37.2	51.9	29.0
I3D-R50	37.5	52.5	29.2	40.9	57.2	31.8
MViT	40.9	57.4	31.7	43.0	57.8	35.4
GLC	43.1	57.0	34.7	44.8	61.2	35.3
CSTS	43.7	58.0	35.1	-	-	-
Ours	**46.52**	**59.72**	**38.09**	**48.39**	**64.61**	**38.67**

Table 7: Comparison with state-of-the-art methods on Ego4D and EGTEA Gaze++ dataset. To save space, we use abbreviations: Rec for Recall, Prec for Precision.

Acknowledgments

This work was support in part by National Natural Science Foundation of China (No. 62301121, No. U23A20286),the Fundamental Research Funds for the Central Universities ZYGX2025XJ002, China Postdoctoral Science Foundation 2023M740529, 2023TQ0046.

References

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; and Feichtenhofer, C. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6824–6835.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19012.

Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19383–19400.

Huang, Y.; Cai, M.; Li, Z.; Lu, F.; and Sato, Y. 2020. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29: 7795–7806.

Huang, Y.; Cai, M.; Li, Z.; and Sato, Y. 2018. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, 754–769.

Huang, Y.; Chen, G.; Xu, J.; Zhang, M.; Yang, L.; Pei, B.; Zhang, H.; Dong, L.; Wang, Y.; Wang, L.; et al. 2024. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22072–22086.

Jacobs, R.; Jordan, M.; Nowlan, S.; and Hinton, G. 1991. [a]Adaptive Mixtures of Local Experts, [o] Neural Computation, vol. 3.

Jiang, Y.; Li, X.; Chen, Y.; He, Y.; Xu, Q.; Yang, Z.; Cao, X.; and Huang, Q. 2022. Maxmatch: Semi-supervised learning with worst-case consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5970–5987.

Jiang, Y.; Xu, Q.; Yang, Z.; Cao, X.; and Huang, Q. 2019. Dm2c: Deep mixed-modal clustering. *Advances in Neural Information Processing Systems*, 32.

Jiang, Y.; Xu, Q.; Zhao, Y.; Yang, Z.; Wen, P.; Cao, X.; and Huang, Q. 2023. Positive-unlabeled learning with label distribution alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15345–15363.

Lai, B.; Liu, M.; Ryan, F.; and Rehg, J. M. 2024a. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, 132(3): 854–871.

Lai, B.; Ryan, F.; Jia, W.; Liu, M.; and Rehg, J. M. 2024b. Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*, 192–210. Springer.

Li, Y.; Fathi, A.; and Rehg, J. M. 2013. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, 3216–3223.

Li, Y.; Liu, M.; and Rehg, J. M. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, 619–635.

Li, Y.; Nagarajan, T.; Xiong, B.; and Grauman, K. 2021. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6943–6953.

Li, Y.-M.; Huang, W.-J.; Wang, A.-L.; Zeng, L.-A.; Meng, J.-K.; and Zheng, W.-S. 2024. Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In *European Conference on Computer Vision*, 363–382. Springer.

Liu, G.; Tang, H.; Latapie, H. M.; Corso, J. J.; and Yan, Y. 2021. Cross-view exocentric to egocentric video synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, 974–982.

Liu, J.-W.; Mao, W.; Xu, Z.; Keppo, J.; and Shou, M. Z. 2024. Exocentric-to-egocentric video generation. *Advances in Neural Information Processing Systems*, 37: 136149–136172.

Ohkawa, T.; Yagi, T.; Nishimura, T.; Furuta, R.; Hashimoto, A.; Ushiku, Y.; and Sato, Y. 2023. Exo2egodvc: Dense video captioning of egocentric procedural activities using web instructional videos. *arXiv preprint arXiv:2311.16444*.

Park, J.; Lee, J.; and Sohn, K. 2025. Bootstrap Your Own Views: Masked Ego-Exo Modeling for Fine-grained View-invariant Video Representations. *arXiv preprint arXiv:2503.19706*.

Pavlitska, S.; Hubschneider, C.; Struppek, L.; and Zöllner, J. M. 2023. Sparsely-gated mixture-of-expert layers for cnn interpretability. In *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–10. IEEE.

Qiu, H.; Li, H.; Zhao, T.; Wang, L.; Wu, Q.; and Meng, F. 2022. RefCrowd: Grounding the target in crowd with referring expressions. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4435–4444.

Qiu, H.; Shi, Z.; Wang, L.; Xiong, H.; Li, X.; and Li, H. 2025. EgoMe: A New Dataset and Challenge for Following Me via Egocentric View in Real World. *arXiv preprint arXiv:2501.19061*.

Qiu, H.; Wang, L.; Zhao, T.; Meng, F.; Wu, Q.; and Li, H. 2024. MCCE-REC: MLLM-Driven Cross-Modal Contrastive Entropy Model for Zero-Shot Referring Expression

Comprehension. *IEEE Transactions on Circuits and Systems for Video Technology*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rahmani, H.; and Mian, A. 2015. Learning a non-linear knowledge transfer model for cross-view action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2458–2466.

Shi, H.; Dao, S. D.; and Cai, J. 2025. LLMFormer: Large language model for open-vocabulary semantic segmentation. *International Journal of Computer Vision*, 133(2): 742–759.

Shi, H.; Hayat, M.; and Cai, J. 2024. Unified open-vocabulary dense visual prediction. *IEEE Transactions on Multimedia*, 26: 8704–8716.

Sigurdsson, G. A.; Gupta, A.; Schmid, C.; Farhadi, A.; and Alahari, K. 2018. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, 7396–7404.

Tavakoli, H. R.; Rahtu, E.; Kannala, J.; and Borji, A. 2019. Digging deeper into egocentric gaze prediction. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 273–282. IEEE.

Thakur, S. K.; Beyan, C.; Morerio, P.; and Del Bue, A. 2021. Predicting gaze from egocentric social interaction videos and imu data. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 717–722.

Wang, Q.; Zhao, L.; Yuan, L.; Liu, T.; and Peng, X. 2023. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3307–3317.

Xu, J.; Huang, Y.; Hou, J.; Chen, G.; Zhang, Y.; Feng, R.; and Xie, W. 2024. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13525–13536.

Xue, Z. S.; and Grauman, K. 2023. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36: 53688–53710.

Yu, H.; Cai, M.; Liu, Y.; and Lu, F. 2019. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1358–1366.

Zhang, M.; Teck Ma, K.; Hwee Lim, J.; Zhao, Q.; and Feng, J. 2017. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4372–4381.